DATA EDITING AT THE NATIONAL CENTER FOR HEALTH STATISTICS

Kenneth W. Harris, National Center for Health Statistics
6525 Belcrest Road, Rm. 915, Hyattsville, MD 20782

Key Words: Survey, Environment, Data Processing, Imputation

Background

The National Center for Health Statistics (NCHS) is the Federal agency responsible for the collection and dissemination of the nation's vital and health statistics. To carry out its mission, NCHS conducts a wide range of annual, periodic, and longitudinal sample surveys and administers the national vital statistics registration systems. These sample surveys and registration systems form four families of data systems: vital event registration systems, population based surveys, provider based surveys, and followup/followback surveys.

Much of what happens to the data covered by these data systems, from collection through publication, depends on the family to which they belong. At most steps along the way, various activities and operations are implemented with the goal of making the data as accurate as possible. These activities and operations are generally categorized under the rubric, "data editing." In the 1990 Statistical Policy Working Paper 18, "Data Editing in Federal Statistical Agencies," [1] data editing is defined as:

> Procedure(s) designed and used for detecting erroneous and/or questionable survey data (survey response data or identification type data) with the goal of correcting (manually and/or via electronic means) as much of the erroneous data (not necessarily all of the questioned data) as possible, usually prior to data imputation and summary procedures. [However, this report <u>includes</u> data imputation procedures.]

As will be shown in this report, data editing procedures vary greatly between NCHS data systems.

Twenty-four data systems, identified in the following section, are included in this report. For each data system, summary descriptions of NCHS data editing practices are provided in the following 11 areas:

- Environment in Which Survey Takes Place
- Data Processing Environment and Dispersion of the Work
- Audit Trail
- Micro-, Macro-, and Statistical Editing
- Prioritizing of Edits
- Imputation Procedures
- Editing and Imputation Standards
- Costs of Editing
- Role of Subject Matter Specialists
- Measures of Variation
- Current and Future Research

Within each of these areas, data editing practices are grouped according to the type of data system, i.e., vital event registration systems, population based surveys, provider based surveys, and followup/followback surveys.

Environment in Which Survey Takes Place

Registration Systems

The vital event registration systems cover six vital events: **Mortality, Fetal Mortality, Abortion, Natality, Marriage and Divorce**. For each of these systems, data are obtained from certificates and reports filed in state registration offices and registration offices of selected cities and other areas. Coverage for each registration system is limited to its prescribed registration area (RA). The oldest registration areas, mortality, fetal mortality, and natality, have been complete since 1933. These three are national data systems; i.e., they cover the entire United States. The marriage RA started in 1957 with 30 states and reached its current coverage of 42 states plus selected areas in 1986. The Divorce RA started in 1958 with 14 states and by 1986 had expanded to 31 states plus selected areas [2,3]. The Abortion RA started in 1977 with five states and reached its current coverage of 14 states in 1987.

Mortality (approximately 2,000,000 annual events), fetal mortality (60,000) and natality (4,000,000) registration are required by all states; registration completeness for the mortality and natality systems exceeds 99 percent. The Abortion RA collects information on approximately 300,000 abortions per year, about 22 percent of the annual U.S. total. (Because of budgetary constraints, NCHS has not processed abortion data since 1993). The Marriage RA, excluding Puerto Rico and the Virgin Islands, covers approximately 81 percent (785,000) of U.S. marriages. The Divorce RA, excluding the Virgin Islands, accounts for 49 percent (280,000) of the annual U.S. divorce count.

In addition to these six registration systems, two other data systems, the Current Mortality Sample **(CMS)** and the Linked Birth and Infant Death Data Set **(LBIDDS)**, are based on data obtained from the Mortality and Natality Registration Systems. The CMS is a 10 percent systematic sample taken from the regular mortality file on a monthly (month of death) basis. The CMS covers the 50 states, the District of Columbia and New York City; it includes 17,000-20,000 deaths per month. The Linked Birth and Infant Death Data Set, which also covers the 50 states, the District of Columbia and New York City, links the more detailed information from the birth certificate with the information from the death certificate for each of the approximately 40,000 infants who dies before his/her first birthday.

Population Based Surveys

Three of the Center's data systems are classified as population based surveys. They are the National Health Interview Survey, National Health and Nutrition Examination Survey, and the National Survey of Family Growth. The designs of these surveys are based on stratified multistage samples of households, where the household is defined as the basic sample unit. Based on established criteria, a person (one or more) in the sample household is selected as the ultimate sample unit, i.e., the unit of analysis.

National Health Interview Survey (NHIS)

The **NHIS** is a continuing nationwide sample survey in which data are collected on the incidence of acute illness and injuries, the prevalence of chronic conditions and impairments, the extent of disability, the utilization of health care services, and other health related topics. Generally, personal interviews are completed in 47,000 households for about 123,000 sample persons.

National Health and Nutrition Examination Survey (NHANES)

The **NHANES** obtains nationally representative information on the health and nutritional status of the American population through a combination of personal interviews (mostly in the respondent's home) and detailed physical examinations. These examinations are conducted in specially equipped mobile examination centers (MEC) that travel around the country. The last survey, NHANES III, the sixth in the cycle of health examination surveys conducted since 1960 [4], collected data on topics such as high blood pressure, blood cholesterol, infectious diseases, diabetes, HIV infection, blood lead levels, allergies, osteoporosis, and other nutritional status measures.

The NHANES III [5], conducted over two 3 year phases, 1988-91 and 1991-94, covered the U.S. civilian, noninstitutional population aged 2 months and older. Each phase constituted a national sample of about 20,000 persons, with an expected interview completion rate of 85-90 percent and a response rate of about 75-80 percent for the medical examination. More than 78 percent of the persons selected for the 1988-91 phase participated in the medical examination. Selected subpopulations, children (< 5 years), older persons (60+), Black Americans and Mexican Americans, were oversampled.

National Survey of Family Growth (NSFG)

The Center's third population based survey, the **NSFG**, is a periodic nationally representative household survey of women of reproductive age (15-44 years). The survey, first conducted in 1973 [6], collects data on fertility and infertility, family planning, and related aspects of maternal and infant health. The 1988 survey, the fourth in the cycle [7], selected 10,000 eligible sample households from the frame of households that participated in the NHIS between 1985 through 1987. A total of 8,450 women were interviewed in person, in their own homes, by trained female interviewers.

Provider Based Surveys

Seven NCHS data systems form the family of provider based surveys, collectively called the National Health Care Survey (NHCS). Included here are the National Hospital Discharge Survey (NHDS), National Survey of Ambulatory Surgery (NSAS), National Ambulatory Medical Care Survey (NAMCS), National Hospital Ambulatory Medical Care Survey (NHAMCS), National Nursing Home Survey (NNHS), National Home and Hospice Care Survey (NHHCS), and the National Health Provider Inventory (NHPI). Whereas population based surveys use the household as the basic sample unit, provider based surveys use the medical provider (physician, hospital, nursing home, etc.) as the basic sample unit. The provider furnishes information on samples of provider/patient contacts, e.g., office visits, hospital stays, nursing home stays, etc.

Samples for these surveys range in size from the approximately 475 emergency rooms in the NHAMCS to the 87,000 facilities covered by the NHPI.

Followup/Followback Surveys

Six of the NCHS data systems included in this report are classified as Followup/Followback surveys. They are:

- National Maternal and Infant Health Survey **(NMIHS)**
- 1991 Longitudinal Followup **(LF)** to the National Maternal and Infant Health Survey
- National Mortality Followback Survey **(NMFS)**
- National Health and Nutrition Examination Survey Epidemiologic Followup Study **(NHEFS)**
- Longitudinal Study of Aging **(LSOA)**

- National Nursing Home Survey Followup (NNHSF)

Sample sizes range from 7,500 to 26,000 persons.

Data Processing Environment and Dispersion of the Work

There are many similarities in the data processing activities employed by NCHS offices for their respective data systems. This is especially true for data systems within the same "family of surveys." For example, registration areas provide NCHS with coded and edited computer tapes or microfilm copies of vital event certificates which are converted to uniform codes and subjected to machine edits. The other surveys use CAPI (Computer Assisted Personal Interview), preliminary hand edits, machine edits, etc. There are, however, a number of procedures that cross "family survey" lines that are gaining greater usage with the rapid advances made in survey technology. Of particular interest to NCHS is "source point data editing" (SPDE). This refers to editing survey data by any means of access to either the interviewer (or other data collector), the respondent, or records within a limited time following the original interview or data collection. The time limit reflects the period within which the persons involved can reasonably be expected to remember details of the specific interview or, in the case of data collected from records, a time within which there is reasonable expectation that there has been no change to the records which would affect the data collected. Thus, data completion and accuracy are much more likely to result when source point data editing is used.

Audit Trail - This term refers to a process of maintaining, either by paper or electronically, an accounting of all changes of sample or survey data item values and the reasons for those changes. The level of effort varies by data systems; some are manual, while others are automated.

Micro-, Macro-, and Statistical Editing

This section describes three types of editing processes. The following definitions are used in this section.

Micro-editing - Editing done at the record or questionnaire level.

Macro-editing - Editing to detect individual errors by checking on aggregated data or by applying checks to the complete set of records.

Statistical editing - Editing based on statistical analysis of respondent data. It may incorporate cross-record checks, as well as historical data.

Micro-, macro, and statistical editing for the eight registration systems are all very similar. Automated edits are designed to (1) assure code validity for each variable and (2) verify codes or code combinations which are considered either impossible or unlikely occurrences.

For each of the other three types of data systems, most or all of the following procedures are used:

- Extensive machine micro-editing.
- Where appropriate, comparison of current estimates with previous years.
- Assuring reasonableness of record counts, sampling rates, etc.
- Checking ranges, skip patterns, consistency of data from different sources.
- Checking medical data for compatibility with age and/or sex.

Priority of Edits

None of the registration systems gives special priority to any item in the editing procedures. The other data systems prioritize their edits based on:

- Identifiers needed to link data files.
- Questionnaire items used to weight sample data to national estimates.
- Medical data incompatible with demographic data.

Imputation Procedures

Imputation is defined as a process for entering a value for a specific data item where the response is missing or unusable.

Registration Systems

Except for Abortion Registration, which does not impute for missing items, imputation procedures among registration systems apply primarily to demographic items. In Mortality registration, imputation procedures are done by machine, which checks for invalid codes. The following variables are subject to imputation procedures: age, sex, date of death, marital status of decedent, race of decedent, and education of decedent.

Missing natality data that are imputed include child's race, sex, date of birth, and plurality. Data imputed for the mother include race, age, marital status and residence. Imputation is done by machine, either on the basis of a previous record with similar information for other items on the record (e.g., mother's age imputed on the basis of a previous record with the same race and total-birth order), or on the basis of other information on the certificate (e.g., marital status on the basis of mother's and father's names, or lack of name). The tape documentation includes flags to indicate when imputation was performed.

Finally, marriage and divorce data imputation are limited to month of marriage (or divorce) and age of bride and/or groom (marriage only). Hot deck and cold deck imputation procedures are used. In hot deck imputation, a missing data item is assigned the value from a preceding record in the same survey having similar (defined) characteristics. In cold deck imputation, a missing data item is assigned the value from a similar record in a previous similar survey.

Population Based Surveys

Imputation procedures for the Center's other surveys differ from those used by the Registration Systems. In the case of the NHIS, unit nonresponse (missing sample cases) is imputed by inflating the sample case weight by the reciprocal of the response rate at the final stage of sample selection, or by a poststratification adjustment based on independent estimates of the population size in 60 age-race-sex categories.

Item non-response (missing question answers) is imputed, where possible, by inferring a certain or probable value from existing information for the respondent. For example, in the NHIS, a missing "husband's age" (or "date of birth") is assigned the value of "wife's age + 2 years."

In the NHANES, the calculation of sample weights addresses the unit nonresponse aspects of the survey except for special cases.

In the NSFG, the sample weights adjust for unit nonresponse. Imputation of missing items in the NSFG was carried out by the contractor. For the most part, a hot-deck procedure was used to impute missing values.

Provider Based Surveys

The provider based surveys have established imputation procedures for three types of nonresponse: unit nonresponse, record nonresponse, and item nonresponse. Unit nonresponse is imputed by inflating the sample weight of similar responding units. Record nonresponse is imputed by inflating the sample weight of similar responding cases to account for the missing cases. Item nonresponse is imputed by inferring a certain or probable value from existing respondent information.

Followup/Followback Surveys

Four of the followback surveys, the LSOA, the NMFS, NHEFS, and the NNHSF, did not impute any data, although missing data items were filled in by using logical relationships as described in the above example. Unknown or inconsistent data were coded as "unknown." The other two used “hot deck” procedures.

Editing and Imputation Standards

For each of its registration systems, NCHS monitors the quality of demographic and medical data on tapes received from the states by independent verification of a sample of records of data entry errors. In addition, there is verification of coding at the state level before NCHS receives the data. All other systems employ error tolerance standards established for interviewer performance (if applicable), and enforced by editing and telephone reinterviews. Error tolerance standards are also established for coding and keying of data, and are enforced by sample verification.

Costs of Editing

The costs of editing are very difficult to determine, though some surveys and data systems appear to have a better handle on this than others.

None of the eight registration systems could provide an estimate of their editing costs. All other systems estimated their data editing costs between 10-30 percent of total survey costs.

Role of Subject Matter Specialists

For all surveys and data systems, the primary role of subject matter specialists is to write edit specifications, from which edit programs are prepared; to review results of edit runs and to adjudicate failures in collaboration with programmers. Their secondary role is to compare standard sets of estimates with historical series to identify anomalies. In addition, they also consult with survey design staff on field edits.

Measures of Variation

- No sampling error for 100 percent registration systems; however estimates of variation are computed for vital events <20.
- Marriage and divorce data tables list sampling error by area expressed as a percent of the area total.
- Other surveys produce estimates of sampling (but not non-sampling) errors.
- Selected surveys present estimates based on assumptions regarding the probability distribution of the sampling error.

Current and Future Research

There are several ongoing research activities and a number of others are being considered for the future.

Resource constraints, both money and personnel, are the major limiting factors. The following represent programmatic changes in data collection, data processing/editing, data analysis, etc., that will occur or be investigated in future years. Aside from these specifics, however, perhaps the biggest change, one which is well underway now, and cuts across all surveys, is the shift from paper and pencil data collection to computerized data collection. This shift makes it more difficult to omit data items, to enter inconsistent or impossible data, etc.

- Implementation of electronic birth and death certificates by the states.
- Implementation of the Super MICAR (Mortality Medical Indexing, Classification, and Retrieval) system by all states. The intent of Super MICAR is to allow data entry operators to enter cause-of-death information as it is literally reported on the death certificate. Under the current MICAR system, cause of death information must be entered using abbreviations or standardized nonmeclature [8]. Implementation of Super MICAR is essential to a successful electronic death certificate system.

Determination of optimum imputation techniques (single and multiple procedures) and their applicability to NHANES.

Evaluation of automated data collection methodology for the NHDS.

Feasibility of developing an automated system for coding and classifying medical entities using the ICD-10.

Feasibility of developing an automated system for data correction and creation of an audit trail (Followback surveys).

Summary

Data editing practices at NCHS are quite extensive. Unfortunately, detailed descriptions of these practices for this report were precluded because of space limitations. However, some summary findings on NCHS data editing practices are provided below and in the following table.

- Among NCHS data systems, the cost of data editing is the least documented variable. Only five data systems provided dollar and other resource costs of their data editing procedures. Most of the other data systems offered "guestimates" of 10-20 percent of total survey costs, with a few "guestimating" as high as 30 percent of total survey costs.
- About sixty percent of the Center's data systems collect data on an on-going basis throughout the year and publish data on an annual basis.
- Two-thirds of the Center's data systems report item non-response rates under 5 percent.
- One third of the Center's data systems use Computer-assisted telephone interviewing (CATI) as their primary or secondary data collection method.
- One-half of the Center's data systems release micro-data with identifiable imputed data items; another one-quarter release micro-data without identifying imputed data items.
- Virtually all NCHS data systems have rules establishing minimum standards of reliability that must be met in order to disseminate data.
- Slightly more than one-third of the Center's data systems monitor analysts/clerks in their data editing procedures; three-fourths monitor their automated editing procedures. However, only three data systems formally evaluate their data editing systems.

References

1. Statistical Policy Working Paper 18: "Data Editing in Federal Statistical Agencies," Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC, May 1990.
2. National Center for Health Statistics: "Data Systems of the National Center for Health Statistics," Vital and Health Statistics, Series 1-No. 16, Hyattsville, MD, December 1981.
3. National Center for Health Statistics: "Data Systems of the National Center for Health Statistics," Vital and Health Statistics, Series 1-No. 23, Hyattsville, MD, March 1989.
4. National Center for Health Statistics: "Cycle I of the Health Examination Survey, Sample and Response, United States, 1960-62," Vital and Health Statistics, Series 11- No. 1, Rockville, MD, May 1965.
5. National Center for Health Statistics: "Sample Design: Third National Health and Nutrition Examination Survey," Vital and Health Statistics, Series 2-No. 113, Hyattsville, MD, September 1992.
6. National Center for Health Statistics: "National Survey of Family Growth, Cycle I," Vital and Health Statistics, Series 2-No. 76, Rockville, MD, June 1977.
7. National Center for Health Statistics: "National Survey of Family Growth, Cycle IV, Evaluation of Linked Design," Vital and Health Statistics, Series 2-No. 117, Hyattsville, MD, July 1993.
8. Harris, Kenneth W.; Rosenberg, Harry M.; et al. "Evaluation of an Automated Multiple Cause of Death Coding System." Proceedings of American Statistical Association, Social Statistics Section, Washington, DC, August 1993.

Frequency of Selected Data Editing Practices Among NCHS Data Systems

	Yes	No	NA/DK[1]
1. Data dissemination limited by confidentiality (privacy) restrictions?	24		
2. Does Data system release microdata (respondent level data)?	19	5	
3. Are imputed items identified?	13	11	
4. For aggregated data, is information provided on percentage of a particular item which has been imputed?	5	16	3
5. Are there minimum standards for reliability of disseminated data?	21	1	2
6. Is information available on the cost of data editing?	5	19	
7. Are there procedures for monitoring editors, clerks, analysts, etc.?	9	14	1
8. Are there procedures for monitoring automated editing procedures?	18	4	2
9. Is there an audit trail (i.e., a record kept) for some or all data editing transactions?	21	2	1
10. Are performance statistics maintained in order to evaluate the data editing system?	3	21	
11. Has any analysis been done on the effect of data editing on estimates produced?	5	19	
12. Is survey data editing information released?	15	9	
13. Is validation editing performed?	22	2	
14. Is macro-editing used?	17	7	
15. Are any other data editing techniques performed?	6	17	1

[1] Not Applicable/Don't Know